Disease variant prediction with deep generative models of evolutionary data

Year: 2021 | Citations: 624 | Authors: J. Frazer, Pascal Notin, M. Dias, Aidan N. Gomez, Joseph K. Min

---

Abstract

Quantifying the pathogenicity of protein variants in human disease-related genes would have a marked effect on clinical decisions, yet the overwhelming majority (over 98%) of these variants still have unknown consequences1–3. In principle, computational methods could support the large-scale interpretation of genetic variants. However, state-of-the-art methods4–10 have relied on training machine learning models on known disease labels. As these labels are sparse, biased and of variable quality, the resulting models have been considered insufficiently reliable11. Here we propose an approach that leverages deep generative models to predict variant pathogenicity without relying on labels. By modelling the distribution of sequence variation across organisms, we implicitly capture constraints on the protein sequences that maintain fitness. Our model EVE (evolutionary model of variant effect) not only outperforms computational approaches that rely on labelled data but also performs on par with, if not better than, predictions from high-throughput experiments, which are increasingly used as evidence for variant classification12–16. We predict the pathogenicity of more than 36 million variants across 3,219 disease genes and provide evidence for the classification of more than 256,000 variants of unknown significance. Our work suggests that models of evolutionary information can provide valuable independent evidence for variant interpretation that will be widely useful in research and clinical settings. A new computational method, EVE, classifies human genetic variants in disease genes using deep generative models trained solely on evolutionary sequences.